PROTECTION OF PERSONAL DATA IN CLINICAL DOCUMENTS - A MODEL APPROACH
## CONTENTS

1. Background .......................................................... 3
2. Introduction .................................................................. 3
3. Privacy Considerations and Scope .................................. 6
4. Anonymization Strategy Challenges ................................... 7
   » 4.1 Anonymization Definitions ........................................ 7
   » 4.2 Anonymization Challenges .......................................... 7
5. Anonymization Approach ............................................... 10
   » 5.1 Assessment of Anonymization Approach ............................. 11
   » 5.1.1 Identifiers ............................................................ 12
   » 5.1.1.1 Direct Identifiers .................................................. 12
   » 5.1.1.2 Indirect (or Quasi) Identifiers .................................... 13
   » 5.1.1.3 Options for Anonymization ..................................... 13
     - Option 1 ...................................................................... 13
     - Option 2 ...................................................................... 14
     - Quantitative Method ...................................................... 14
     - Qualitative Method ....................................................... 15
   » 5.2 High-Risk Sections within Clinical Documents .................... 16
     - 5.2.1 Narratives ............................................................ 16
     - 5.2.2 Listings .............................................................. 19
     - 5.2.3 Investigators, Authors, and Other Study Personnel ........... 19
     - 5.2.4 Images ............................................................... 19
   » 5.3 Anonymization Report ............................................... 19
6. Technical Considerations or Options ................................. 21
References ............................................................... 23
Appendix 1: Terms and Definitions ....................................... 25
   » 2.1: Anonymization Methodology ........................................ 28
   » 2.2: Identification of Data Variables (Direct and Indirect Identifiers) .......... 28
   » 2.3: Assessment of Anonymization ...................................... 29
     - 2.3.1: Fulfilment of the Criteria for Anonymization ..................... 29
     - 2.3.2: Risk Assessment .................................................... 29
   » 2.4: Data Utility Considerations ......................................... 31
   » 2.5 Conclusion ........................................................... 31
Appendix 2: Anonymization Template .................................... 28
   » 2.1: Anonymization Methodology ........................................ 28
   » 2.2: Identification of Data Variables (Direct and Indirect Identifiers) .......... 28
   » 2.3: Assessment of Anonymization ...................................... 29
     - 2.3.1: Fulfilment of the Criteria for Anonymization ..................... 29
     - 2.3.2: Risk Assessment .................................................... 29
   » 2.4: Data Utility Considerations ......................................... 31
   » 2.5 Conclusion ........................................................... 31
Appendix 3: Example of Anonymization Approaches for Coded Medical Terms and Regions ...................................... 32
List of Abbreviations .................................................... 33
CLINICAL STUDY REPORTS APPROACH TO PROTECTION OF PERSONAL DATA

1. Background*

TransCelerate BioPharma Inc. is a non-profit organization of biopharmaceutical companies focused on advancing innovation in research and development (R&D), identifying and solving common R&D challenges, thus increasing the quality of clinical studies and delivering more high-quality medicines to patients. Accordingly, TransCelerate has undertaken a commitment to enhance public health and medical and scientific knowledge by addressing sharing and transparency of clinical trial information. Compliance with applicable national, regional, and local privacy and data protection laws is not optional.

2. Introduction

The regulatory environment continues to evolve with regard to clinical trial data and document sharing. In response to the changing landscape (Figure 1).

In July 2013, Pharmaceutical Researchers and Manufacturers of America (PhRMA) and European Federation of Pharmaceutical Industries and Associations (EFPIA) member companies demonstrated a commitment to sharing and making transparent clinical trial data by publishing “Principles for Responsible Clinical Trial Data Sharing”. This included principles to enhance public access to clinical study information and Clinical Study Reports (CSRs) as follows: For any submissions filed as of January 1, 2014, following approval of a new medicine or new indication for an approved medicine in the United States (US) and/or European Union (EU), biopharmaceutical companies will make publicly available, at a minimum, the synopses of CSRs for clinical trials submitted to Food and Drug Administration (FDA), European Medicines Agency (EMA), or national competent authorities of EU Member States. PhRMA and EFPIA member companies also agreed to evaluate requests from qualified scientific and medical researchers for access to full CSRs including participant-level clinical trial data, study-level clinical trial data, and protocols from clinical trials in participants for medicines and indications approved in the US and the EU as necessary for conducting legitimate research. These transparency principles likely have applicability in major markets other than the US and EU as privacy considerations are not company-specific but are region- or country-specific in the diverse global privacy landscape.

*Nothing in this paper should be construed as legal advice, nor does anything in this paper imply or warrant that use of this approach complies with applicable laws or regulations. Users implement the approach outlined in this paper at their own risk and bear the sole responsibility for ensuring their compliance with applicable laws and regulations in their respective jurisdictions.
In September 2014, TransCelerate published “Clinical Study Reports Approach to Protection of Personal Data” providing a general approach that can be applied across most clinical documents and reports, with the caveat that adjustments may be needed to comply with local national privacy laws and regulations. This paper was published to offer implementation options to fulfill PhRMA and EFPIA commitments. At the time, tools to accomplish privacy protection were limited; “redaction only” was primarily used for documents, whereas other anonymization techniques were used for data. Redaction remains the current methodology used by EMA for protecting privacy when Access to Documents requests are made through Policy 0043.

Redaction is also the typical method of protection for company proprietary and confidential information in requests from all sources and would be the method of choice for other document types that fall under the EU Clinical Trial Regulation of 2014, like Inspection Reports and Serious Breach Reports.

In October 2014, the EMA published its policy on the publication of clinical data, referred to as Policy 0070, along with a Q&A guide. Under the Phase 1 of the implementation of Policy 0070, the EMA publish “clinical reports” defined in the Policy as clinical overviews, clinical summaries and CSRs included in the dossier of Marketing Authorization Applications (MAAs). Policy 0070 applies to new MAAs and Article 58 applications submitted via the centralized marketing authorization procedure, as of 01 January 2015, and extension of indication and line extension applications relating to existing centrally authorized medicinal products submitted as of 01 July 2015. The general policy was followed by the publication in March 2016 by EMA of a comprehensive guidance which provides details to assist sponsors with implementation of Policy 0070. The publication of Individual Patient Data (IPD) is not in scope of this current EMA guidance but will be the subject of future guidance.

In April 2015, TransCelerate published “Data De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach.” Its primary focus was to address the de-identification and anonymization of Individual Patient Data (IPD) to meet transparency and disclosure commitments and respond to research requests while safeguarding the privacy of individuals (eg, participants and company staff). This paper highlights which techniques are currently available to conform to existing directives and regulatory guidance, while balancing the utility of the de-identified data to the researcher. The scope of this paper assumes that there is a legally-binding data sharing agreement between the data provider and the researcher prior to sharing data, and that the data provider has defined a secure method for sharing de-identified and anonymized data. Sharing of data outside of these assumptions, eg, public sharing, is out of scope of this version of the paper. The IPD paper better clarifies the data release context and expands sections considering the Expert Determination method, quantification of risk, and technology needs.
This document, “Protection of Personal Data in Clinical Documents – A Model Approach,” is an update of “Clinical Study Reports Approach to Protection of Personal Data”\(^5\) that reflects the EMA's Policy 0070 guidance issued in March 2016 to support the publication of clinical data.\(^10\) It represents current knowledge, experience, and technology available so far. This paper could be updated in the future as experience is gained and the understanding of technology in this area advances. The original paper published in September 2014 may still be applicable for guidance in connection with sharing clinical documents. For example, Appendix 3: Summary of Approach by CSR section is an illustrative implementation aid, which will need to be refined according to individual sponsor CSR practices\(^5\). Redaction practices may also be useful for other documents as mentioned in the Clinical Trials Registration and Results Information Submission as published in Federal Register\(^13\). Protocols and Statistical Analysis Plans will have to be made available under this new rule.

The revised paper outlines points for consideration to manage the risk of re-identification when clinical documents are shared publicly for secondary uses once the concerned procedure has been completed: the medicinal product is registered, withdrawn, or withdrawn due to a negative opinion. When determining their anonymization strategy, sponsors should consider global implications and legal and regulatory requirements to protect the privacy of individuals, groups and staff associated with a clinical study.

The protection of Commercially Confidential Information (CCI) is out of scope for this paper; however, several guidances issued by the EMA\(^8, 10, 14, 15\) discuss this topic.

**Figure 1: TransCelerate Papers and Emerging Regulatory Landscape**

\(^*\) Clinical Study Reports Approach to Protection of Personal Data

\(^**\) Data De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach
3. Privacy Considerations and Scope
Privacy concerns are paramount in the context of public disclosure of clinical documents.

In general, the following should be considered when sharing clinical documents:

» A model and consistent approach seeks to protect privacy, regardless of the audience.

» Sponsors will need to assess the risk of re-identification of a study participant when determining the anonymization approach, including the risk that data may be linked together to other data sources.

» Sponsors may consider taking a risk-based approach when balancing the extent to which the content of a clinical document will be anonymized while still striving to maintain an acceptable level of data utility (see Section 5.1.1.2 on Quantitative and Qualitative methods for more information). Any such approach will need to include a consideration of applicable data protection laws and privacy laws, in particular in the EU.

» Particular care should be taken, and a more cautious approach adopted to disclosure, when study or participant factors are present that increase the risk of re-identification (eg, rare diseases, small or vulnerable populations, or low-frequency events).

» Aggregated data or descriptions of aggregated data and study-level information (eg, public register IDs, tabular, graphic, or cross-participant data) may be disclosed publicly if sponsors determine that the risk of re-identifying individuals represented in the data are unlikely to raise privacy concerns and could be disclosed publicly. Exceptions that may require anonymization of aggregated data include analyses by indirect identifiers (eg, tables by country/age/gender).

In all cases, the sponsor should ensure that individual study participant (patient) privacy is protected.

Anonymized datasets and CCI are out of scope for the model approach described in this paper.
4. Anonymization Strategy Challenges

In addition to the previous channels of disclosure, EMA Policy 0070 (October 2014) and EMA Guidance for Implementation (March 2016) provide sponsors with expectations and implementation details for making clinical documents in the scope of the Policy publicly available on the EMA website. Policy 0070 implementation guidance emphasizes the crucial importance of protecting individuals’ privacy while seeking to maintain utility of data for public use, which can present a real challenge for sponsors (in particular due to continuous technological developments, eg, in data mining and data linking). Due consideration must be given to ensure the protection of the privacy of personal information associated with a clinical document while minimizing the impact on the utility of the data. (See details on data utility, sometimes referred to as scientific or clinical utility).

4.1 Anonymization Definitions

Different usages and meanings of terminology have hindered a common understanding of privacy protection techniques. Some sources use the terms “de-identification” and “anonymization” interchangeably. Others use “de-identification” to describe a process and “anonymization” to denote a specific kind of de-identification that cannot be reversed (ie, the key is destroyed). This paper will adopt the definitions used by the International Organization for Standardization (ISO) as well as the EMA: “Anonymization is the process of turning data into a form that does not identify individuals and where identification is not likely to take place. It allows for a much wider use of the information.”6, 8, 9, 16 As discussed above, EU data protection law contains specific definitions for some of these terms (including anonymous data), which in some cases differ to those used in this paper.

Anonymization of trial information may be performed using one (or more) of several potential methods: Information can be removed, redacted/masked, or replaced with generalizations or transformations that allow analysis. Details on each of these techniques are found in Section 5.0.

4.2 Anonymization Challenges

There are differing opinions on how to assess re-identification risk.17 These may be compounded in the future especially as public databases for cross-linking become more extensively available. No process provides a guarantee of zero risk, but instead offers a methodology for reducing risk. Future approaches may call for writing clinical documents with broad distribution to the public in mind. Establishing a reasonable anonymization approach or a combination of approaches is a challenge, including the need to remain consistent across documents.
Data utility will often be a key consideration when selecting the anonymization methodology. If anonymization of the data results in clinical documents that are no longer useful for their intended secondary purposes, data utility is compromised. Furthermore, anonymization of clinical documents results in the data being altered in some way. Ensuring that the analysis results produced after anonymization are similar to the results that would be obtained from the original clinical document is critical. Therefore, the amount of distortion should be minimized. Ultimately, EMA has recommended that a balance must be reached to obtain an acceptably low probability of re-identification, to protect study participants and satisfy the requirements of data protection laws, and sufficient data utility.\(^\text{10}\)

Finally, in implementing EMA Policy 0070, sponsors discover many initial submissions for which Policy 0070 is applicable include documents completed often many years before EMA Policy 0070 Guidance was issued (March 2016). In preparing these “legacy” documents to be made public, redaction techniques (e.g., ‘masking’) may be the only viable option to protect privacy. Moreover, considering the pros and cons of alternative means of anonymization and the challenges in implementing them, redaction techniques may also continue to be applied to newly created clinical documents for some time based on individual sponsor consideration. These redaction techniques, while protecting privacy, could limit data utility. Participant safety narratives pose a particular challenge for sponsors as these narratives often contain personal information or identifiers which, if retained, could jeopardize privacy by providing potential re-identification “data links” to trial participants. (Section 5.2.1)

It may be that, for some studies, there is no practical way to sufficiently anonymize the data to protect the participants, while still maintaining the utility of the data. In these circumstances, the requirements of data protection law may mean that certain data cannot be disclosed without some other lawful basis being found (e.g., obtaining specific participant consent as is expected in a journal publication).

The examples of challenges outlined above underscore the conclusion that anonymization approaches are not simple checklist exercises. Tools, rules, and algorithms require customization to specific trials and diseases. Human intervention and assessment must supplement the tools.

To address these challenges, one option for sponsors to proactively consider public disclosure while authoring clinical documents. In other words, taking into account the potential audience – e.g., the general public, not simply a governmental agency – when drafting clinical documents. Indeed, one might consider proactive authoring as a “privacy by design” approach to address privacy and mitigate re-identification risk, while minimizing the need for further anonymization. Of course, with this new
approach comes a new mindset and considerations:

» Proactive Authoring:
  - Will this approach, which minimizes the potential for re-identification and protects personal privacy, still adequately support regulatory review and decision-making?

» Anonymized document:
  - Will the rules around redaction, removal, generating replacement text or transforming text allow for utility of the data?
  - What level of anonymization of study participant safety narratives will provide protection and maintain usefulness?

The sponsor’s role is to make thoughtful, well-informed decisions about document anonymization procedures that are in compliance with global data protection regulations and still adhere to the EMA Policy 0070 stipulations. The aim of this paper is to provide sponsors with points to consider during decision-making. By sharing the considerations from the collective experiences across TransCelerate member companies and other stakeholders, this paper aims to help sponsors develop knowledge and a framework for their decision-making so they can more efficiently navigate the requirements of transparency, privacy protection, and the desire to maintain data utility for multiple audiences and purposes.
5. Anonymization Approach

In the context of this paper, anonymized health data are data that are sufficiently devoid of personal information to protect personal privacy and meet the requirements of data protection and privacy laws. In order to ensure study participants are not likely to be re-identified by disclosed clinical documents, sponsors must consider the various approaches available and the residual risk – that is, the risk that remains after controls are taken into account – that may allow for the re-identification of a study participant. Some of the options include removal, redacting/masking, and replacement (see Appendix 1 for definitions).

The assessment of what information to anonymize and by what method(s) is based upon the determination of the potential risk of re-identification of a study participant by disclosing the information. Consideration should be given to direct and indirect personal identifiers suggested by the Health Insurance Portability and Accountability Act (HIPAA)\(^\text{18}\) and EU Data Protection Regulations No. 45/2001\(^\text{19}\) and Data Protection Directive 95/46/EC,\(^\text{20}\) (soon to be replaced by the EU General Data Protection Regulation 2016/679), which define personal data broadly to encompass any information relating to an identified or identifiable natural person. Note that the EU data protection law is broader in scope than the US concept of “personally identifiable information” (or “PII”), and the protection of personal data is a fundamental right of EU citizens. “Personal data” under EU law means any data which relates to an identified or identifiable individual. To determine whether a study participant is identifiable from the data, account should be taken of all the means reasonably likely to be used to identify him/her. This includes considering the available technology, the amount of time and effort it would take to re-identify the data, and the costs involved.

Also it is useful to consider the minimum standard for de-identifying data described in well-known articles and secondary sources, such as Hrynaszkiewicz.\(^\text{27}\)

Each method has advantages and disadvantages that must be considered. For instance, removal and redaction both drastically reduce the chance of re-identification of a person and can – at least partly – be automated; however, data utility of the reports is significantly diminished. While dates and ages may be replaced, or generalized by use of ranges, substitution of other indirect identifiers (eg, sex, race, and medical terms) would need to consider how the analysis is affected (except for the examples provided in Appendix 3 proposing the replacement of coded medical terms). An advantage of format-preserving replacement is that a reader cannot distinguish between real and transformed data when a value is missed by the anonymization technique (ie, hiding in plain sight).\(^\text{22}\) Any method that replaces data with anonymized data creates a seemingly analyzable document, however, potential risk exists that a person still could be re-identified if not all identifiable data has been obscured within a document or that the replaced content may lead to incorrect conclusions or lead to uncertainty. In addition, controls must be put in place to avoid confusing an original
document with one that is anonymized. For example, sponsors should be alert to identify unique pharmacovigilance safety cases from anonymized reports so as not to double count safety cases that may be reported in literature. Hence, the importance of the anonymization report is to inform researchers of the techniques applied to the clinical document to avoid inconsistent or incorrect conclusions.

When transforming data, there is a need to consider the integrity of the final product. For example, if participant’s age is anonymized using randomization, this may result in an inappropriate conclusion on the effect of age, thereby distorting the original report. Thus, it is important to remember that implementation of anonymization methods could present unintended consequences such as altering results and interpretations of data within the clinical documents.

Per EMA Policy 0070, there are modules of the clinical study report that are out of scope such as investigator curriculum vitae (CVs)/biographies, and signatures and therefore these can be removed from the document prior to submission to the EMA.

5.1 Assessment of Anonymization Approach
Sponsors will be faced with developing an anonymization approach taking into account global regulatory requirements as appropriate. The following parameters should be considered:

» The level of anonymization required;
» What methodologies to use;
» The clinical utility of the data once shared/posted

Sponsors must consider the risk threshold level when determining how to implement an anonymization approach. Multiple factors can be used to determine what level of risk of re-identification a sponsor may be willing to assume and the resulting degree of anonymization required to protect privacy. For example:

» Data use agreements with researchers accessing the documents serve to decrease the risk of re-identification compared with public web postings with minimum privacy controls.
» Small trial population size and rare diseases usually present a higher risk of re-identification compared with very large studies in common disease areas.
» Placing an individual study in the context of an entire marketing application may add to the re-identification risk due to the possibility of relating information between documents.

To select the most appropriate method, it is necessary to have a clear understanding of the data in the clinical documents (ie, all direct and indirect identifiers within the document, see Appendix 1 for definitions). 22, 23
5.1.1 Identifiers
Perhaps the most important category of data impacted by anonymization are identifier data. Moreover, how identifier data are processed during anonymization will depend on whether they are categorized as direct or indirect identifiers.

Direct identifiers are data that directly identify a single individual without additional information or with cross-linking through other information that is in the public domain – what might be termed “obvious” identifiers. Some examples of direct identifiers are the individual’s name, address, social security number or other identifying number or code, telephone number, email address, images, or biometric record, according to ISO/TS 25237:2008\(^\text{17, 23}\).

By contrast, indirect (or quasi) identifiers are data that can identify a single person only when used together with other indirectly identifying data. Indirect identifiers can reduce the population to which the person belongs, possibly down to one if used in combination. Some examples of indirect identifiers are sex, age, and date of birth according to ISO/TS 25237:2008.\(^\text{17, 23}\) Note that under EU data protection law, indirect identifiers will likely still be considered “personal data”, where there are means which are reasonably likely to be used to re-identify the data.

5.1.1.1. Direct Identifiers
To protect against the risk of re-identification of a study participant, it is generally advisable to redact all direct identifiers within clinical documents except for study participant IDs, which can be replaced through pseudonymization\(^\text{22}\) instead of redaction as shown in Table 1. Using the same replacement study participant ID allows identifying the same subject across a document and does make the data more useful. Furthermore, direct identifiers are not analyzed.

<table>
<thead>
<tr>
<th>Anonymization Approach</th>
<th>Example text</th>
<th>Risk to re-identify*</th>
<th>Linkability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Text</td>
<td>Study participant 018462 experienced an AE of headache........Study participant 018462 withdrew early from the study due to the AE of headache.</td>
<td>High</td>
<td>able to follow a participant</td>
</tr>
<tr>
<td>Redacting Text</td>
<td>Study participant [redacted] experienced an AE of headache........Study participant [redacted] withdrew early from the study due to the AE of headache.</td>
<td>Low</td>
<td>unable to follow a participant</td>
</tr>
<tr>
<td>Replacement Text (with the same random ID at all instances)(^\text{16, 17})</td>
<td>Study participant [redacted] experienced an AE of headache........Study participant [redacted] withdrew early from the study due to the AE of headache.</td>
<td>Moderate</td>
<td>able to follow a participant as in the original text</td>
</tr>
<tr>
<td>Replacement Text (with different random ID at each instance)</td>
<td>Study participant [redacted] experienced an AE of headache........Study participant [redacted] withdrew early from the study due to the AE of headache.</td>
<td>Low</td>
<td>unable to follow a participant</td>
</tr>
</tbody>
</table>

* depends on other unique indirect identifiers

Note: Red denotes text changed as a result of the anonymization approach chosen
5.1.1.2 Indirect (or Quasi) Identifiers
Indirect identifiers present a greater challenge for sponsors than direct identifiers. Sponsors should carefully analyze what data within a particular dataset is an indirect identifier, taking into account the potential risk of re-identification and the utility of the documents. Sponsors should carefully analyze the various anonymization methods and select the most appropriate method. Whereas direct identifiers are removed from the data without reducing its utility, indirect identifiers often convey some sort of information that can be useful for analysis and so removing them is more likely to impact on the utility of the data. In some cases, however, this may nonetheless be necessary in order to protect individual privacy.

5.1.1.3 Options for Anonymization
In the case of EMA Policy 0070 submissions, EMA states that a sponsor can establish the adequate anonymization of clinical documents by 1 of 2 options: (1) the fulfilment of the criteria for anonymization (see Appendix 2.3.1 for details) or whenever a proposal does not meet these criteria, (2) an evaluation of the risk of re-identification (see Appendix 2.3.2 for details) with clear justification of why a particular anonymization method was selected. To describe the method selected the Marketing Authorization Holder (MAH) is required to submit an Anonymization Report with the submission documents that will be posted on the EMA website (See Section 5.3).

Option 1
Option 1, stipulated in the Anonymization Report, is a set of 3 criteria used to evaluate the anonymity of the document as set forth by the Article 29 Data Protection Working Party:

- No possibility of singling out an individual;
- No possibility to link records to an individual;
- Information concerning an individual cannot be inferred.

According to EMA guidance, if all the criteria for Option 1 have been fulfilled, sponsors do not need to complete Option 2. However, given the limited experience applying this option, it is difficult to provide guidance around implementation until more is understood. It is clear that it has, so far, been very difficult to retain useful data and still meet these criteria and therefore, few, if any, sponsors have adopted this approach.

Option 2
As noted in the Anonymization template, there are 2 methods currently available for the assessment of risk in clinical documents:

- Quantitative: A method that analyzes the data itself to measure the risk and how to best de-identify the data by establishing a probability. The approach measures the risk as a numerical value.
» **Qualitative**: A method that assesses the risk of re-identification based on the characteristics of the source data (Figure 2). The approach uses a qualitative scale (eg, high, moderate, low) for the assessment of risk.

To ensure the highest data utility of clinical documents, sponsors should develop a systematic process to analyze indirect identifiers within clinical documents and anonymize according to the appropriate risk level.

See Appendix 2.3.2 for additional details on these 2 approaches for the assessment of risk.

**Quantitative Method**
A pre-defined risk threshold of re-identification should be established and justified. Currently the EMA has referenced a risk threshold of 0.09 as acceptable and cited in literature (See Institute of Medicine Report Appendix 2).

This risk threshold is based on assessments of healthcare data and may need to be evaluated for applicability to clinical trial data. As the number of published submissions increases, a greater understanding of the development and application of threshold factors through specific work examples will become available from which we can learn from and refine our common understanding. Over time there might be other thresholds considered to represent a suitable risk for clinical data in submissions.

The level of de-identification should increase until the risk of re-identification is lower than the set threshold. There are a number of factors that sponsors can consider when developing an algorithm to assess risk, such as:

» The number and quality of direct and indirect identifiers within a report or record;
» The size and nature of the population for the disease studied (eg, is the disease rare or common, pediatric population?);
» The size of the study (eg number of study participants);
» The number of study centers and their distribution across countries (eg, global vs a single country);
» The size of the groups of study participants that share the same indirect identifier values (equivalence classes) which is the basis of k-anonymity (Refer to Appendix 1 Terms and Definitions)

Limited commercial solutions with the capacity to provide an analytical risk assessment of clinical documents are available for sponsors to consider. See Appendix 2.3.2 for additional details on the quantitative method.

**Qualitative Method**
This method assesses the risk of re-identification on a qualitative scale. Factors to be considered are largely the same as for the quantitative method.
As illustrated in Figure 2, given a particular condition based on the “rarity of disease”, “number of patients”, “number of study sites”, a risk level could be derived and the appropriate “Anonymization Approach” could be adopted.

**Figure 2. Illustration of a Qualitative Approach to Anonymization Based on Risk Level**

<table>
<thead>
<tr>
<th>HIGH RISK</th>
<th>MODERATE RISK</th>
<th>LOW RISK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rarity of patient population</td>
<td>Remove/anonymize direct and indirect identifiers</td>
<td></td>
</tr>
<tr>
<td>Number of patients in the study</td>
<td>Remove/anonymize direct and KEY indirect identifiers</td>
<td></td>
</tr>
<tr>
<td>Number of study sites in the study</td>
<td>Remove/anonymize only direct identifiers</td>
<td></td>
</tr>
</tbody>
</table>

*Refer to Table 2 for more detail

In general, the risk of re-identification increases with small studies or studies with only a few sites. Similarly, rare patient populations may also increase the risk of re-identification. Special care is needed when assessing the level of anonymization chosen for these studies due to the high level of personally identifiable information. These factors should be considered carefully by the sponsor when determining anonymization approach.

See Appendix 2.3.2 for additional details on the qualitative method.

5.2 **High-Risk Sections within Clinical Documents**

After completing the general risk determination for a study, sponsors should specifically decide the most appropriate anonymization technique for high-risk sections of a clinical document, which contain a large amount of personal information and, therefore, carry a high risk of re-identification. For retrospective applications, when data analysis and document writing are complete, redaction may be the most practicable option. For proactive applications, when data analysis and document writing are ongoing, there are alternative options in addition to redaction options available (eg, data replacement).
5.2.1 Narratives

Study participant narratives pose a challenge for protecting privacy. Narratives inherently include large amounts of personal participant information, specifically direct identifiers and many indirect identifiers that could be used to re-identify an individual, especially when correlating data across documents and with other data sources. It is imperative that sponsors consider the content of the narratives as a whole and select an anonymization option that minimizes the risk to study participant’s privacy. Individual studies may require different approaches according to the risk level as detailed above in Figure 2.

Table 2 below provides an illustrative, incremental approach to the differing levels of anonymization that could be employed to ensure risk level (as defined in Figure 2) is mitigated appropriately for purposes of EMA Policy 0070:

<table>
<thead>
<tr>
<th>Study Risk Level</th>
<th>Anonymization Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOW</td>
<td>Option A: Anonymization of direct identifiers (study participant ID) and the following limited indirect identifiers: - Demographic information - Study participant-associated dates - Geographic location when associated with a participant - Medical history - Verbatim text</td>
</tr>
<tr>
<td>MODERATE</td>
<td>Option B: Anonymization of direct identifiers, indirect identifiers in Option A, and the following: - Event outcomes - Diagnostic test results</td>
</tr>
<tr>
<td>HIGH</td>
<td>Option C - Additional redaction inclusive of Option B: Exact details to be redacted will be determined on a case-by-case basis</td>
</tr>
</tbody>
</table>

There are multiple factors a sponsor must consider while determining the appropriate risk mitigation strategy. These include (but are not limited to):

» Options A and B include anonymization of both direct and indirect identifiers. Where possible incremental additional redaction could be automated.

» For Option B, technology can be used to identify data that are in the document (eg. diagnostic test results or finding text string); however, anonymization of textual information is more challenging (event outcomes or diagnostic results).

» Option C, case-by-case evaluation is required to reach a balance between protecting privacy and providing data utility in order to determine what to anonymize in those circumstances of high risk.
Alternatively, a step-wise approach could be used to determine how much should be anonymized to appropriately mitigate the risk. For this approach, identifiers are categorized as Level 1 (identifiers that will usually be stable over time, eg, geographic location) and Level 2 (indirect identifiers that include assessments, data points captured throughout the clinical trial\textsuperscript{23}.

Other considerations:

» If text such as medical terminology is anonymized there is a risk that medical content is changed (see Coded Terms for Rare Adverse Events).

» Replacing dates and ages can easily be automated when anonymizing data sets; however, in clinical documents accommodating the new text within the original text spacing and formatting is more challenging and could lead to misinterpretation.

» Keep in mind, replacement and resynthesizing in a text section has numerous complications that do not exist when anonymizing a data set.

Furthermore, it is recommended to adapt the approach chosen for each study or submission, taking into consideration any special circumstances such as deaths due to rare events or sensitive data such as alcoholism, drug abuse, any risky behavior, or history of sexually transmitted.

Verbatim Text
Verbatim text can contain unique information that can lead to study participant re-identification (ie, unstructured terms that are not coded) and will likely require additional scrutiny. Table 2 describes options that can be applied to verbatim text.

Coded Terms for Adverse Events
In general, retaining coded terms helps to preserve data utility. However coded terms could significantly contribute to the risk of re-identification and hence should be included in the risk assessment. Careful consideration should be given when selecting the anonymization methodology factoring in the level of risk and data utility. See examples in Appendix 3.

Dates
Dates are generally considered indirect identifiers; however, when considering the likelihood of re-identification of a person, birth dates and death dates are more specific identifiers than assessment dates and can be related or linked more easily to external databases adding to risk. Different approaches may need to be considered for the various dates within a document. Some options include redacting day/month only, off-setting dates by a set factor, or applying a relative study day method to dates. Converting dates into relative ages and age banding or aggregating data into a single
category such as “age 90 or older” is a way to avoid giving specific participant dates. The applied method(s) should be consistent within a document, to avoid incorrect conclusions.

More information on these methods in TransCelerate’s paper, “Data De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach,” Regulatory requirements for the disclosure of dates associated with adverse event reports and study participant narratives should be considered when assessing an anonymization method for dates. Proactively establishing a set method for the presentation of dates in a generalized fashion eg, as relative days, within future original clinical documents may alleviate the need for further anonymization of those data.

**Geographical Locations**

Geographic locations should also be considered when assessing the criteria for anonymization of the narrative section of a clinical document. Currently, there is no generally accepted threshold for what population size will limit the risk of re-identification. However, some possible considerations for sponsors include:

» Have geographical locations been factored into the analysis in light of the disease incidence and prevalence?
» Are there global sites or is the study at a single site?
» How many countries are included in the study?
» Are there regional or ethnic differences among study participants that could be identified?
» What is the threshold set for disclosure (ie, city, state, region, country, or continent)?
» Can the regions/countries data be aggregated to minimize the risk of re-identification?

5.2.2 **Listings**

It is important to note that while by-patient/by-visit listings are out of scope of Policy 0070 for Phase 1, EMA consider listings where parameters or outcomes values for selected patients and selected visits or a certain single time point (eg, Sections 14.3.1 Adverse Events, 14.3.2 Listings of Deaths, Other Serious and Significant, and 14.3.3 Narratives) are within scope of Phase 1 of Policy 0070 and must be included in the submission to EMA and risk of re-identification minimized. All patient/by visit listings should be removed with an agreed textual explanation added, as specified by EMA.

5.2.3 **Investigators, Authors, and Other Study Personnel**

As a general rule, EMA asserts that privacy and data protection laws allow the release of the name(s) of the sponsor(s) and principal or coordinating investigator(s) within clinical documents. Names of subinvestigators can be redacted or anonymized. It is
important to note that handwritten signatures are also personally identifiable data and therefore should be fully redacted. Names of experts or designated personnel with legally defined responsibilities and roles with respect to aspects of the Marketing Authorization dossier (e.g., Qualified Person, Qualified Person for Pharmacovigilance, Clinical expert, Investigator) are included in the dossier because they have a legally defined role or responsibility and it may be in the public interest to release these data. As discussed above, all EU member states have specific legislation and/or specific national rules, guidelines or practices on the protection of personal data and therefore, in these countries, these data may need to be redacted. Contact details (e.g., personal address, personal phone number) should not be disclosed. Additionally, other personnel (e.g., document authors, monitors, external partners, etc.) should be redacted.

**5.2.4 Images**
Care should be taken when considering the disclosure of images within a clinical document. The process of anonymizing an image can be challenging due to the data that may include hidden personally identifiable information.

As anonymization technology evolves, sponsors should adapt their processes from removal of images to anonymization, if possible.

**5.3 Anonymization Report**
The EMA Policy 0070 requires the submission of an anonymization report. The anonymization report’s main objectives are to explain what changes have been made to the clinical documents included in a marketing authorization application to the EMA in order to protect study participants’ privacy; to provide the rationale for those changes; and to demonstrate that after anonymization, the risk of re-identification is at an acceptable level and the impact on data utility has been considered.

The needs of two audiences must be considered:

First, EMA reviewers will evaluate the anonymization report to determine whether the sponsor has exercised due diligence to reduce the risk of re-identification to an acceptable level. EMA guidance is careful to state that the agency has prescribed no particular anonymization methodology—that choice is left to the applicant—but reviewers do expect to see evidence that some kind of logical, evidence-based process has been put into place, based on thoughtful assumptions. Moreover, EMA reviewers need to understand the applicant’s stated rules with enough clarity to permit them to assess, by inspection of the anonymized clinical documents, whether the applicant has executed the planned transformations systematically and consistently.

Second, members of the public will rely on the anonymization report to understand what elements of the published clinical documents have been altered from the version reviewed by regulatory authorities. Furthermore, some members of the public may
subject an applicant’s methods to scrutiny: eg, to determine whether a pharmaceutical company has appropriately balanced the needs of study participant’s privacy and data utility. In this respect, the anonymization report is an opportunity to share a straightforward and open discussion of the considerations the applicant weighed during the anonymization process, and any particular challenges, which may enhance public understanding.

Neither audience may find particular benefit in an exhaustively technical document. Thus, to the extent possible, the anonymization report should be concise and clear. Publications that support the applicant’s decision-making should be cross-referenced where appropriate. The EMA has provided a template in the External Guidance\textsuperscript{10} which should be suitable for most submissions. It can be subdivided into sections covering different report types if needed (see Appendix 2 for template and more details).
6. Technical Considerations or Options

Historically, anonymization of clinical documents was a labor-intensive, manual process largely in response to specific requests for data from health authorities or payers. With the implementation of EMA Policy 0070, the volume of clinical documents to be anonymized will grow substantially, resulting in a reliance by industry upon electronic tools to anonymize personal data and redact CCI within these documents.

A number of tools exist that allow simple redaction of text using a blue/black bar to hide the text. These tools originated to support other industries and legal, government, and healthcare documents. However, most lack any artificial intelligence specific to clinical documents that will be made public under Policy 0070.

The EMA has acknowledged that no mature technology is widely available yet to support sophisticated anonymization of the massive number of clinical documents that will be available for public disclosure. They acknowledge that further consultation with stakeholders is needed to achieve the full implementation of Policy 0070 and desired data utility over time. Today, transparency principles are shaping the requirements for a new set of tools that must come to the market to enable success in delivering data utility while protecting study participant’s privacy.

Given the lack of viable tools able to support compliance with Policy 0070, all stakeholders should actively encourage new tool development that will provide the following:

- Measure risk quantitatively and provide a real risk measurement of the probability of re-identification based on the rules applied to the document;
- Apply anonymization techniques to common document formats (eg, Adobe PDF documents) beyond the ability to redact, such as replacing with analyzable values, randomization, generalization, resynthesis;
- Support the EMA implementation and review cycles as outlined in the procedural guidances from 02 March 2016;
- Reliability, scalability, affordability, and general commercial readiness to be used broadly;
- Use of artificial intelligence/machine-learning for anonymization;
- Other innovative solutions.

Yet even when capable anonymization tools become widespread, the approach to compliance with Policy 0070 will likely be a combination of using an automated technology (a tool) as well as a manual process in order to properly anonymize
clinical documents prior to submission to the EMA. Some sponsors have successfully
developed automated processes (eg. SAS™ macros) to be used for de-identification of
individual participant-level data sets, and there are currently efforts to extend these
rules and logic to apply to CSRs as well.

The appropriate combination of process and technology will vary based on the tool
chosen and the resulting process required to support it. Some considerations include:

» Who owns the process – is it centralized or on a per-submission team basis?
» Who is involved in the review process?
» How are co-development partners integrated?
» Is the tool purchased to use in-house or hosted?
» Will the vendor who owns the tool perform the work as a service?
» Does the tool use multiple anonymization techniques or simply redaction?
» Who is accountable for the anonymization report, and how will it be made specific to each
package?
» How is risk being managed – qualitative or quantitative measures?
» How will the redaction proposal be turned into the final redacted package? What tools will
be used and who will do this?
» How will CCI be identified and when, in the process, will it be marked within the package?

Elements that go beyond tools and processes that sponsors should consider include:

» How will all the various public iterations of study results align across registries, manuscripts,
posters, EMA website, plain language summaries, documents received via Freedom of
Information Act / Access to Documents and requests by researchers/investigators?
» How can sponsors confidently reduce the risk of re-identification with the increasing volume
of publicly available personal information in social media and other developing public
databases?

Answers to these questions will start to shape a process and help organizations
understand where tools will be required within their document delivery process.
TransCelerate plans to monitor the technical landscape and communicate changes
as needed. An ongoing assessment to evaluate the emerging technical capability to
quantify the risk of re-identification in anonymized documents would be worthwhile.
References


Appendix 1: Terms and Definitions

The terminology in use in data sharing and data protection can be confusing; some terms are used interchangeably, others are in greater use in a particular geographic region. For clarity, the list below provides the definitions as used in this paper as well as some terms as defined by EMA. Note that these definitions are not used consistently in global privacy laws.

anonymization
The process that removes the association between the identifying data and the data subject. In anonymized data, the data subject cannot be identified by the recipient of the information.\(^{17}\)

clinical data
Clinical documents and individual patient data (IPD) as defined in the EMA Policy 0070 Implementation Guidance.

clinical documents/reports
The clinical overviews (submitted in module 2.5), clinical summaries (submitted in module 2.7), and the clinical study reports (submitted in module 5, “CSR”) together with the following appendices to the CSRs: 16.1.1 (protocol and protocol amendments), 16.1.2 (sample case report form), and 16.1.9 (documentation of statistical methods).

commercially confidential information (CCI)
Any information contained in the clinical documents which is not in the public domain or publicly available and where disclosure may undermine the legitimate economic interest of the company.

data subject/study participant
An individual who is the subject of personal data; persons to whom data refer (ISO/TS 25237:2008).

de-identification
The process of rendering data into a form that does not identify individuals and where identification is not likely to take place.

So, a general term for any process of removing the association between a set of identifying data and the data subject (ISO/TS 25237-2008).\(^{17,24}\)

direct identifier
A piece of data that can be used to uniquely identify an individual (eg, study participant ID, social security number, exact address, telephone number, email address, government assigned identifier) without additional information or cross-linking other information that is in the public domain.\(^{17,26}\)

disclosure
The act of making data available to one or more third parties.

final redacted document package
A “Final Redacted Document” package shall contain the final redacted versions of all clinical documents related to one single, finalized regulatory procedure that falls under the scope of EMA Policy 0070.

final redacted version
This is the clinical document version, submitted by the applicant/Marketing Authorization Holder (MAH) for publication, which should reflect the EMA review outcome (accepted/rejected redactions).

identifiable person
One who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural, or social identity\(^{17}\) (ISO/TS 25237:2008).

indirect identifier/quasi identifier
Data, which in connection with other information, can be used to identify an individual with high probability, eg, age at baseline, race, gender, events, specific findings, etc.

» a quasi-identifier Level 1 is not likely to change over time, is visible, and is available in other sources. Typically, it is demographic data such as gender, age at baseline, country, BMI, etc.;

» a quasi-identifier Level 2 is longitudinal information that is likely to change over time such as measurements, events, etc.\(^{21}\)

individual patient or participant data (IPD)
The person-specific data separately recorded for each participant in a clinical study.

k-anonymity
A technique “to release person-specific data such that the ability to link to other information using a quasi-identifier is limited.”\(^{27}\)
K-anonymity is a framework developed by Sweeney for quantifying the amount of manipulation required of the quasi-identifiers to achieve a given desired level of privacy. The technique is based on the concept of an equivalence class, the set of records that match on all quasi-identifier values. A data set is said to be k-anonymous if, for every combination of quasi identifiers, there are at least k-matching records. For example, if a data set that contains the quasi identifiers birth year and state, has k=4 anonymity, then there are at least four records for every combination of birth year and state.

**masking**
The process of concealing or covering identifying data irreversibly from view, generally with a black bar or box. Legally, redaction is also used and is meant to prevent the inadvertent disclosure of privileged or otherwise confidential or sensitive information by the permanent removal of visible text and images from documents. See also redaction

**personally identifiable information (PII)**
Any information about an individual maintained by an agency, including but not limited to, education, financial transactions, medical history, and criminal or employment history and information which can be used to distinguish or trace an individual’s identity such as their name, social security number, date and place of birth, mother’s maiden name, biometric records, etc., including any other personal information which is linked or linkable to an individual. (SP 800-122).

**protected personal data (PPD)**
From Directive 95/46/EC: “Personal data” shall mean any information relating to an identified or identifiable natural person (‘data subject’); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural, or social identity.

**Pseudonymization**
A privacy preservation technique that both replaces the direct association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms. Typically, pseudonymization is implemented by replacing direct identifiers (like the subject’s name) with a pseudonym, such as a randomly generated value. (ISO/TS 25237:2008). Pseudonymization allows linking information belonging to an individual across multiple data records or information systems, provided that all direct identifiers are systematically pseudonymized. The technique is recognized as a method for privacy protection of personal health information. It can be performed with or without the possibility of re-identifying the subject of the data (reversible or irreversible pseudonymization). 16,17

**Redaction**
Redaction is the separation of disclosable from non-disclosable information by blocking out individual words, sentences, or paragraphs or the removal of whole pages or sections prior to the release of the document. 26

See also masking

**redaction proposal document package**
The “Redaction Proposal Document” package shall contain the redaction proposal versions of all clinical documents related to one, single, finalized regulatory procedure that falls under the scope of Policy 0070, along with a number of additional documents listed in the “External guidance on the procedural aspects related to the submission of clinical documents for the purpose of publication in accordance with EMA Policy 0070.” 10

**redaction proposal version**
The clinical document version containing the applicant’s/Marketing Authorization Holder’s (MAH) proposed redactions of CCI and personal data. These proposed redactions should be highlighted in a ‘read-through’ manner. 10

**re-identification**
The process of analyzing data or combining it with other data with the result that individuals become identifiable, sometimes also referred to as ‘de-anonymization.’

**re-identification risk**
The risk that de-identified records can be re-identified. Re-identification risk is typically reported as the percentage of records in a data set that can be re-identified. 16 (IOM report, Appendix B) 25

**residual risk**
The risk that remains after controls are taken into account (the net risk or risk after controls). 16 (IOM report, Appendix B) 25

**risk**
The probability of re-identifying a trial participant. 16 (IOM report, Appendix B) 25

**sensitive data**
Any data that, in the event of re-identification, would harm a study participant in terms of employability, reputation, insurability, self-esteem or results in loss of income. Examples include history of alcoholism, drug abuse, risky behavior or history of venereal disease.
Appendix 2: Anonymization Template

As described in Section 5, an anonymization report is required for each Policy 0070 submission. The suggested headings below are excerpts of the template provided in the EMA guidance as areas of consideration.

2.1: Anonymization Methodology

Explain how the initial risk of re-identification of study participants was calculated or estimated using the data that formed the basis of clinical documents, or from the clinical documents themselves. What assumptions were made? What variables were considered? In a qualitative approach (non-analytical), one could provide these results in a descriptive fashion (“more frequent”, “higher risk”, “lower risk”) and give justification, referencing published literature if available. In a quantitative approach (analytics), one might provide frequency counts for the different identifiers or indirect identifiers found in the documents, and a numeric estimate of the risk posed by different variable types.

What de-identification criteria were employed to make decisions that reduced the risk of identification to an acceptably small level? On what standards were the methods based? Include references to relevant sources. To justify reliance on a particular approach, one might consider these general principles:

- Is there empiric evidence to support that the methods used provide anonymization, based on prior experience? (qualitative approach)
- Is there a statistical basis to support that the method will reduce the risk of re-identification to a very low level? (quantitative approach)
- Do alternative methods have weaknesses that render them potentially inappropriate in the context of this data release?

2.2: Identification of Data Variables (Direct and Indirect Identifiers)

Provide a list or table that names each type of identifier found in the clinical documents and describes what transformation will be applied (eg, pseudonymized, generalized, offset, retained, redacted, etc.). Existing methods or standards can be cross-referenced (eg, PhUSE standards). If there are identifiers that will not be transformed, explain the reason(s). Explain any considerations that drove the choice of a particular transformation method. Rules for anonymization of a variable can potentially differ among the reports included in a submission or across the sections of a given document. If such context-dependent criteria are set for how a variable will be handled, document this clearly and explain the thought process behind these decisions. A structured list or table addressing different scenarios may be useful to reviewers.
The treatment for a given variable can be dramatically influenced by the assumptions a reviewer makes about the external environment such as the likelihood of data attacks and the information an adversary is likely to know. While these types of assumptions should be laid out explicitly in the “Risk Assessment” section, the “Identification of Data Variable” section should be logically consistent with those assumptions and should cross-refer to them if needed.

2.3: Assessment of Anonymization

There are two options available to establish whether data are adequately anonymized for purposes of EMA Policy 0070. Select an option below (2.3.1 Fulfilment of the Criteria for Anonymization or 2.3.2 Risk Assessment) which applies to the submission. The EMA is clear that only one of these options should be used.

2.3.1. Fulfilment of the Criteria for Anonymization

Explain how the anonymized documents meet each of these three criteria:

» No possibility to single out an individual.
» No possibility to link records relating to an individual.
» No information can be inferred concerning an individual.

Whenever a proposal does not meet one of these criteria, complete Section 2.3.2 instead.

2.3.2. Risk Assessment

First, state any assumptions about which adversary(ies) might conduct an attack (ie, an attempt at re-identification), what their objectives might be, what outside information they might be likely to know, and what methods they are likely to use. Historical precedent may be cited here, but it is also appropriate to state reasonable assumptions about future developments. The potential damage in the event of a successful attack will contribute to the selection of a risk threshold, and these factors should be discussed to justify the choice of a threshold.

The next step is to designate a risk threshold and to evaluate whether the risk of re-identification in the anonymized reports is below the chosen threshold. The content of this section will vary depending on whether your anonymization method is quantitative or qualitative; two different approaches to this section are outlined below.
Quantitative:
Based on the recommendations made in the Institute of Medicine (IOM) report25 and the available precedents for public release of health data, EMA believes that it is advisable to set the threshold to a conservative level of 0.09. Since this level of risk is based on healthcare data, it might be appropriate to assess other risk thresholds for clinical trial data in the future. However, it is up to the applicant to decide on the most appropriate risk scenario and threshold for public disclosure of the clinical document at stake, and if a different threshold is selected a justification shall be provided.

Explain how the residual risk of re-identification was calculated after the clinical documents were anonymized; describe or refer to the algorithms used. Compare the results to the initial risk assessment. If the measurement of residual risk following anonymization of the clinical documents reaches or exceeds the threshold, effective anonymization has been demonstrated.

Qualitative:
Describe the risk threshold used for direct and indirect identifiers. For a qualitative approach, a risk threshold such as “low” or “very low” would be used.

Describe how anonymized reports were assessed against the anonymization criteria which were set up in Section 5.1.1.2 Qualitative Method (eg, through manual review of the reports or the application of automated tools). Compare the result to the selected risk threshold. An argument that the risk of re-identification is acceptably low could be constructed as follows:

» Rules for the anonymization of identifiers in the reports were established based on existing best practices (which should be referenced), which may have been originally derived based on qualitative or quantitative methods. Experience has shown that with these methods, the risk of re-identification is low.

» In the document(s) to be disclosed, where a variable was judged to pose moderate to high risk of re-identification, it was manipulated to reduce the risk to a low level.

» All the remaining variables in the document constitute “low” to “very low” risk according to accepted standards in common use today.

This line of argument relies heavily on empiric evidence based on past experience. It follows that initial anonymization reports will be difficult to justify using this reasoning, as will innovative methods or standards. It may be that only after methods have been substantiated with a quantitative analysis, tested in the real world, and found to provide an acceptable level of anonymization, can they then be integrated into a qualitative approach to anonymization.
2.4. Data Utility Considerations
Data utility refers to the scientific usefulness of the data after it is anonymized, ie, whether the data can still be effectively for the secondary purposes.

Based on the risk methodology chosen, assess and comment on the impact on data utility. Describe any methodology used to measure data utility (eg, expert examination or the precision metric). What thresholds are considered acceptable for the methodology used? From which sources have these thresholds been obtained? Compare findings to the thresholds to assess data utility.

State that the anonymization method was developed with due consideration to the importance of data utility. Where explicit trade-offs were made (eg, redaction vs replacement; manipulating vs retaining certain indirect identifiers which might be important for data analyses), these can be discussed or cross-referenced. Corporate programs are intended to enhance collective scientific knowledge through responsible data sharing with qualified researchers who are best positioned to make use of data to enhance public health.

2.5 Conclusion
Summarize the outcome of the anonymization process and provide a declaration that the anonymization report has been prepared following the guidance made available by EMA, and that the anonymization techniques have been applied consistently in the preparation of the documents comprising the submission to the EMA. Take care not to imply that anonymization per se has been performed according to EMA guidance. EMA’s position is that it is the sponsor’s responsibility to select and implement an anonymization methodology and to ensure their compliance with any applicable national data privacy laws and also include a declaration that the anonymization complies with applicable national data protection legislation. EMA has made clear that the agency does not prescribe or endorse any specific approach.
## Appendix 3: Example of Anonymization Approaches for Coded Medical Terms and Regions

<table>
<thead>
<tr>
<th>Anonymization Type</th>
<th>Example</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>Study participant 310 with history of stomach cancer had an AE of diarrhea. One study participant in Guatemala with history of stomach cancer withdrew due to an AE.</td>
<td>Some re-identification risk from linkability of records; High utility for analyses of AE Preferred Term x outcome; High utility for analyses of tumor type x AE Preferred Term.</td>
</tr>
<tr>
<td>Partial redaction:</td>
<td>Study participant 310 with history of ______ cancer had an AE of ________. One study participant in _______ with history of _______ cancer withdrew due to an AE.</td>
<td>Low re-identification risk from linkability of records; Low utility for analyses of AE Preferred Term x outcome; Low utility for analyses of tumor type x AE Preferred Term.</td>
</tr>
<tr>
<td>Consistent replacement with a higher dictionary coded term:</td>
<td>Study participant 310 with history of malignant and unspecified gastrointestinal neoplasms had an AE of gastrointestinal and motility and defecation conditions. One study participant in Central America with history of malignant and unspecified gastrointestinal neoplasms withdrew due to an AE.</td>
<td>For medical terms we recommend higher level MedDRA terms (eg, the high level group term or the system organ class) rather than preferred terms Lower re-identification risk from linkability of records; Still some utility for analyses of AEs.</td>
</tr>
</tbody>
</table>
List of Abbreviations

Clinical Study Reports (CSRs)
Commercially Confidential Information (CCI)
European Federation of Pharmaceutical Industries and Associations (EFPIA)
European Medicines Agency (EMA)
European Union (EU)
Food and Drug Administration (FDA)
Health Insurance Portability and Accountability Act (HIPAA)
Individual Patient Data (IPD)
Institute of Medicine (IOM)
International Organization for Standardization (ISO)
Marketing Authorization Application (MAA)
Marketing Authorization Holder (MAH)
Personally Identifiable Information (PII)
Pharmaceutical Researchers and Manufacturers of America (PhRMA)
Research and Development (R&D)
United States (US)